



Integrate Web and Standalone BLAST

Combining standalone blast+ with web BLAST to better serve your sequence analysis need

<https://blast.ncbi.nlm.nih.gov/> & <https://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST/>

National Center for Biotechnology Information • National Library of Medicine • National Institutes of Health • Department of Health and Human Services

Introduction

NCBI's web-based BLAST services, freely available from the [BLAST homepage](#), are well-known to the research and education communities. For local batch processing and searching against custom dataset, NCBI also makes the standalone blast+ package available through its [ftp site](#). Less well-known is that web BLAST services and standalone blast+ tools can be integrated with each other in many ways. This handout shows how this integration is useful through some practical examples. These examples below assume the blast+ package is installed and configured properly on a local machine. Refer to the help manual for more details on installation [1].

Integration of Standalone blast+ with Web BLAST Services

The web BLAST services and the standalone blast+ are integrated in the following ways:

- **BLAST searches:** all standalone blast programs, such as blastn and blastp, provide the “-remote” option. This option submits the search to NCBI BLAST server for processing, and saves the returned results to output file in specified format. In addition to the standard set of databases, it also allows the access of other specialized databases. Those specialized databases can be located through an API search against the blastdbinfo Entrez database. More information on is available from an NCBI blogpost [2].
- **Result formats:** web BLAST search results can be requested in different formats. Local blast+ searches can be saved in an archive format (-outfmt 11 option) for reformatting into other formats using the blast_formatter tool. This tool can also format an unexpired web BLAST result using the assigned RID (through -rid VALUE option).
- **Search strategies:** a web BLAST search' settings can be saved to a MyNCBI account as a search strategy for reuse. A saved search strategy can be downloaded for sharing or for use with standalone blast+. Conversely, a search strategy exported from a standalone blast+ search can be used to start a web BLAST search.

Use Case 1: Submit search through standalone blast+ to NCBI BLAST Server

Search programs from the standalone blast+ package have a built-in “-remote” switch that submits a locally constructed search to NCBI BLAST server. Features offered by this switch include:

- Reduced overhead comparing to the web graphical user interface
- More options to fine tune search settings
- Additional customization for tabular output
- Convenient integration with other workflows

The following commandline is an example remote search:

```
blastn -query unknown_seq.fna -db nt -dust yes -evalue 0.05 \
-task dc-megablast -template_type coding_and_optimal -template_length 21 -word_size 11 \
-outfmt "6 delim=; qacc sacc staxid qstart qend sstart send length pident gapopen gaps qcovhsp \
evalue bitscore" -max_target_seqs 5 -out unk_tabular_output.txt -remote
```

The commandline above runs blastn, with the unknown_seq.fna as query (-query unk_seq.fna) against the nt database (-db nt), set expect value of 0.05 (-evalue 0.05), and discontiguous megablast (-task dc-megablast) as the program. It uses two template settings (-template_type coding_end_optimal) and template length of 21 (-template_length 21), a word size of 11 (-word_size 11). It requests a customized tabular output without header (-outfmt “6 ...” more on this field below), asks for top 5 hits (-max_target_seqs 5), and saves the output to the specified file (-out unk_tabular_output.txt). The “-remote” switch specifies that the search is sent to the NCBI web BLAST service.

The -outfmt ‘6 ...’ customizes the tabular output, uses semicolon as delimiter (delim=;) for the space-delimited fields within the quotes:

Field name	qacc	sacc	staxid	qstart	qend	sstart	send	length	pident	gapopen	gaps	qcovhsp	evalue	bitscore
Meaning	Query Accession	Subject Accession	Subject Taxid	Query Start	Query End	Subject Start	Subject End	Alignment Length	Percent Identity	Gap Opening	Gaps	Query Coverage by HSP	Expect Value	Bit Scores

The result can be examined with shell command, such as cat:

```
cat unk_tabular_output.txt
C.USH2A.28.R.TEST;KF493877;10359;1;2273;226458;224186;2273;99.824;0;0;100;0.0;4088
C.USH2A.28.R.TEST;KF493877;10359;510;2273;28865;30628;1764;99.830;0;0;78;0.0;3174
C.USH2A.28.R.TEST;KF493876;10359;1;2273;226378;224106;2273;99.824;0;0;100;0.0;4088
C.USH2A.28.R.TEST;KF493876;10359;510;2273;28866;30629;1764;99.830;0;0;78;0.0;3174
C.USH2A.28.R.TEST;GU980198;10359;1;2273;2273;99.824;0;0;100;0.0;4088
C.USH2A.28.R.TEST;GU980198;10359;510;2273;202755;201012;1773;94.811;4;38;78;0.0;2805
...
```

Use Case 2: Retrieval of search results using blast_formatter tool

Standalone blast+ package's BLAST search programs can save a search result in the archive format (-outfmt 11). The blast_formatter tool can generate other dedicated/specialized formats, such as tabular, XML, or JSON from the archive format. This is a great time and resource saver since it helps avoid re-running the BLAST search to get a different format.

For an example search below saves the result in archive format:

```
blastn -db refseq_rna -query Q.fna -taxids 9823 -outfmt 11 -out test_archive asn
```

The result can be converted into customized tabular format with this command:

```
blast_formatter -archive test_archive asn -outfmt "6 delim=; qacc sacc staxid qstart qend sstart \ send length pident gapopen gaps qcovhsp evalue bitscore"
```

The above commandline Here blast_formatter reads in the archive (-archive test_archive.asn), and requests the customized output without header, with the tabular columns given within quotes. It sets the field separator to semi-colon (delim=;), adds a non-default query coverage per hsp (qcovhsp) field, and prints the result to the console since no output file is specified (no -out argument):

```
U40024;NM_214383;9823;1;7785;1;7785;7785;100.000;0;0;100;0.0;14377  
U40024;NM_214383;9823;1766;2114;1598;1946;351;78.632;4;4;4;1.32e-57;230  
U40024;NM_214383;9823;1598;1946;1766;2114;351;78.632;4;4;4;1.32e-57;230  
U40024;NM_214383;9823;1619;1747;2123;2251;129;82.171;0;0;2;4.98e-22;111
```

The output column specification is the same as used in Use Case 1 (p.1). The complete list is also available from the blast_formatter -help command, under the -outfmt section.

The blast_formatter can also format web BLAST search results, if a valid request ID (RID) is provided through the -rid switch, such as in the example below:

```
blast_formatter -rid 5BTPGY7S013 -outfmt "7 delim=; qacc sacc staxid ssiname qstart qend sstart send \ length pident gapopen gaps qcovhsp evalue bitscore"
```

The format is custom tabular with comment lines (# initialed lines), with columns contain subject taxid and subject scientific name. The latter requires the local installation of the [taxdb files](#).

```
# BLASTN 2.11.0+
# Query: NM_214383.1 Sus scrofa zonadhesin (ZAN), mRNA
# RID: 2GHXGR3S013
# Database: refseq_rna
# Fields: query acc., subject acc., subject tax id, subject sci name, q. start, q. end, s. start, s. end,
alignment length, % identity, gap opens, gaps, % query coverage per hsp, evalue, bit score
# 4 hits found
NM_214383;NM_214383;9823;Sus scrofa;1;7785;1;7785;7785;100.000;0;0;100;0.0;14377
NM_214383;NM_214383;9823;Sus scrofa;1766;2114;1598;1946;351;78.632;4;4;4;1.33e-57;230
NM_214383;NM_214383;9823;Sus scrofa;1598;1946;1766;2114;351;78.632;4;4;4;1.33e-57;230
NM_214383;NM_214383;9823;Sus scrofa;1619;1747;2123;2251;129;82.171;0;0;2;5.01e-22;111
# BLAST processed 1 queries
```

The custom tabular output is the most convenient format for parsing (for structured parsing programmatically, use XML or JSON format). This format can be customized to include other columns such as the title of the matched subjects, their sequences, and their taxonomic name. The following command customizes the output by using customized field delimiter (-delim=;), getting the standard column set (std), and extending it further by appending three new columns to the output, namely subject title (stitle), subject sequence (sseq), and subject scientific name (ssciname):

```
blast_formatter -rid 2GHXGR3S013 -outfmt '7 delim=; std stitle sseq ssciname'
```

```
# BLASTN 2.11.0+
# Query: NM_214383.1 Sus scrofa zonadhesin (ZAN), mRNA
# RID: 2GHXGR3S013
# Database: refseq_rna
# Fields: query acc.ver, subject acc.ver, % identity, alignment length, mismatches, gap opens, q. start, q. end, s. start, s. end, evalue, bit score, subject title, subject seq
# 4 hits found
...
NM_214383.1;NM_214383.1;82.171;129;23;0;1619;1747;2123;2251;5.01e-22;111;Sus scrofa zonadhesin (ZAN),
mRNA;CCACCACCCCACCGAAAGGACCACCCCCACCATAGGACCACCTCCACTGAAAGGACCACCATCCACGAAAAAGACCACTGTTCCCACAGAAA
AAACCATTATCCCCACTGAAAGGACCA;Sus scrofa
# BLAST processed 1 queries
```

This approach could be useful for downloading a large result set from a web BLAST search.

Use Case 3: Exchange search settings between web BLAST and standalone blast+

A web BLAST result page has a “Save Search” link (A) that allows the saving of the search setup permanently to one’s MyNCBI account. Saved settings are accessible through the “Saved Strategies” link (B), which presents available entries in a table (C).

The “download” link (D) saves the search setup to a local file (in the ASN.1 format), which can be shared among colleagues, or be used in a classroom setting to provide a controlled start point during an exercise or exam. The “Choose File” button (E) allows the locating of a strategy file from local disk for upload through the “View” button (F).

Strategies can be shared between standalone blast+ and the web interface.

The following command uses the search strategy saved from one of the listed examples above (C):

```
blastn -import_search_strategy 9NCN8EH8016_strategy asn -remote -out blast_from_strategy.txt
```

The command uses blastn program to read the saved search strategy file (-import_search_strategy 9NCN8EH8016_strategy.asn), enables remote search at NCBI (-remote), and saves the result to a local file (-out blast_from_strategy.txt). Note that there is no input query sequence argument in the command since the query is included in the strategy file. The result will be in the default pairwise format.

Dropping the “remote” option will run the search on the local machine. This requires that the target database is installed locally.

Conversely, a local search setup can be saved to a search strategy file for use with a web BLAST search. This does require that the target database selected is available from the NCBI BLAST server. The follow example command exports the search strategy to a file (relevant switch in red):

```
blastp -db nr -query Q.aa -word_size 6 -seg yes -comp_based_stats 2 -evalue 1e-10 -max_target_seqs 500 \
-outfmt 7 -export_search_strategy SARS2_strategy_standalone.asn -out test_run_tsv.txt
```

Steps to test this:

- Generate a search strategy (or download a precomputed one from above search [here](#))
- Go to [BLAST homepage](#), then the “Saved Strategies” link (G)
- Use “Choose File” button to select the search strategy file from local directory (H)
- Use the “View” button (I) to upload selected strategy file as a preconfigured search page
- Use the “BLAST” button to submit the search

Program	Created	Title	Database	view	download	x
megablast	2021-05-11 12:46:01	(2) - F12249:HSC37F021 normalized infant brain cDNA...	nt	view	download	x
megablast	2021-05-11 12:45:36	F12249:HSC37F021 normalized infant brain cDNA...	nt	view	download	x
megablast	2021-02-03 17:42:15	Nucleotide Sequence	refseq_ma	view	download	x
tblastn	2018-11-09 19:04:45	sp P09601 (288 letters)	WGS_VDB://CAAA01 WGS_VDB://AAHY01 WGS_VDB://AEKR01	view	download	x
megablast	2018-11-09 19:03:18	13 sequences (0 D BACTERIA_1013879)	n/a	view	download	x
blastx	2018-03-14 22:53:46	missed_burk	nr	view	download	x

Use Case 4: Manually construct a PSI-BLAST PSSM in the web BLAST interface and use that to search a custom nucleotide datasets in local tblastn search

The follow functions/features form the foundation for this approach:

- Position-specific iterative blast, i.e., PSI-BLAST, is a more sensitive protein search. The increased sensitivity comes from custom position-specific score matrix (PSSM) constructed at each of the iterations. The PSSM file can be saved for use elsewhere.
- Web PSI-BLAST allows manual adjusting of the hits in between iterations, so undesired hits (such as partial sequences or near identical matches) can be removed.
- The blast+ standalone package provides the tblastn program for finding matches to an input proteins encoded in genomic/transcriptomic assemblies.
- This tblastn program can read in a PSSM generated for the input query protein to perform a more sensitive search.

The following example uses a web PSI-BLAST generated PSSM in a standalone tblastn search to identify potential candidate matches in the genomic and transcriptomic assemblies for a target organism of interest.

Step 1. Run web PSI-BLAST

Open protein BLAST page, load the sequence (Pvull C4-DNA methylase) as the query, select “PSI-BLAST” as the program, open the “algorithm parameters” section and increase the max target seqs to 1000. The configured search page is linked [here](#). Run the search for three iterations, examine the matches after each iteration, remove or add hits as desired.

Step 2. Save the PSSM from the third iteration

In the BLAST result page, click “Download All” and select “PSSM” option at the end. A copy of the precomputed PSSM generated in step 1 is available [here](#).

Step 3. Download and format the WGS and TSA assemblies and convert them into blastable databases

This consists of three tasks: download the desired WGS and TSA sequence datasets from NCBI, inflate the compressed file, and convert the sequence files into blastable databases. In a Linux environment, the tasks can be chained together through the pipe symbol (i.e., the vertical bar character '|').

```
curl 'https://ftp.ncbi.nlm.nih.gov/sra/wgs_aux/PD/FQ/PDFQ01/PDFQ01.1.fsa_nt.gz' | gunzip -c | \
makeblastdb -in - -dbtype nucl -parse_seqids -title "PDFQ01 japonica WGS" -out PDFQ01_wgs
```

This set of commands uses curl to fetch the WGS assembly from the NCBI ftp site and passes it to gunzip to inflate to console (-c). The reconstituted FASTA sequences are passed to makeblastdb for processing:

- -in -, takes input from the console stream instead of an input file
- -dbtype nucl, makes a nucleotide database
- -parse_seqids, indexes the sequence ids to allow specific sequence retrieval by seqids
- -title “PDFQ01 japonica WGS,” provides a title to the database, and
- -out PDFQ01_wgs.txt, names the resulting database files with this base name

Use the same approach to make the TSA blast database:

```
curl 'https://ftp.ncbi.nlm.nih.gov/sra/wgs_aux/GF/YC/GFYC01/GFYC01.1.fsa_nt.gz' | gunzip -c | \
makeblastdb -in - -dbtype nucl -parse_seqids -title "GFCY01 0. japonica TSA" -out GFCY01_tsa
```

Step 4. Run tblastn search with the saved PSSM against the databases created above

Search the PSSM against the databases generated above:

```
tblastn -db PDFQ01_wgs -in_pssm 9WSMVKN016-PSSM_Scoremat.asn -outfmt 7 -out matches_wgs_pssm.txt
tblastn -db GFCY01_tsa -in_pssm 9WSMVKN016-PSSM_Scoremat.asn -outfmt 7 -out matches_tsa_pssm.txt
```

Note the input PSSM through the “-in_pssm 9WSMVKN016-PSSM_Scoremat.asn” call precludes the usage of input query through the “-query file_name” option. The query is in the input PSSM file.

For comparison, do the following search against the same databases using the initial [methylase protein sequence](#) as the query:

```
tblastn -db PDFQ01_wgs -query pvu2_mthase_aa -outfmt 7 -out matches_wgs_query.txt
tblastn -db GFCY01_tsa -query pvu2_mthase_aa -outfmt 7 -out matches_tsa_query.txt
```

Comparing the tblastn PSSM search with the query input counterpart, it is clear that pssm input finds more significant hits. Functionally conserved regions not identified by query input are picked up by the PSSM searches.

References

1. BLAST Help manual. <https://www.ncbi.nlm.nih.gov/books/NBK1762/>
2. Blastdbinfo: API access to a database of BLAST databases. <https://go.usa.gov/x6rct>